UCID- 18665

"Remarks On The Assessment, Representation, Aggregation and Utilization of Expert Opinion"

Terrence L. Fine

April 1980

Lawrence Livermore Laboratory

This work was supported by the United States Nuclear Regulatory Commission under a Memorandum of Understanding with the United States Department of Energy.

Table of Contents

"Remarks On The Assessment, Representation, Aggregation and Utilization of Expert Opinion"

Terrence L. Fine*
School of Electrical Engineering
Cornell University
Ithaca, NY 14850

and

Lawrence Livermore Laboratory
April 1980

ABSTRACT

This report considers the relevance of recent ideas in the foundations of probability to the rational use of expert opinion in the design of a nuclear waste repository, and the assessment of its performance. The main probability concepts introduced are those of modal ('probably A'), comparative ('A is at least as probable as B') and interval-valued ('the lower probability of A is $\underline{P}(A)$ and the upper probability of A is $\overline{P}(A)$') probabilities. We then outline an approach first using comparative probability to model the results of binary elicitation of an expert's opinions concerning repository uncertainties and then employing interval-valued probability to represent comparative probability in a computationally convenient form. We further consider the issue of aggregating or amalgamating the responses of several experts, and we emphasize the need to preserve some measure of the disagreements among the experts. The resulting aggregated interval-valued representation of the responses concerning the uncertainties surrounding the performance of a nuclear waste repository design can then be used to numerically asssess this performance in a manner parallel to that of utility theory. Utility theory is the basis for statistical decision theory. Our recommendations can only be tentative, and research is recommended to gain some working experience with the results of the proposed decision-making process in the repository design context.

*Visiting Professor, Information Systems Laboratory, Durand Building, Stanford University, 1979-1980

I. Introduction

A. The Setting

In the course of summer employment at LLL, I have been exposed to the issues that arise in risk assessment for nuclear waste repositories. The goal is an assessment of the 'harm to man' that can arise from, radioactive waste stored in a repository. 'Harm to man' is assumed to be calculable from knowledge of the emission history of stored radioactive material into the biosphere. This emission is expressible in terms of annual Curie rate reaching the surface in the area of the repository. The scenarios that I have heard considered are confined to radioactive release into surface waters and did not include such catastrophic and rapid possibilities as seismic disturbances uncovering the repository or unwitting human entrance into the repository. The release scenario most spoken of, and the one contemplated in the exercise known as Mock Site A, concerned the seepage of ground water into the repository, the subsequent corrosion of the canister containing solidified waste, the dissolution of the waste into the ambient ground water, and the transportation of this radioactive ground water to an aquifer and eventually to the surface waters. The evaluation of the waste transport even in this 'steady-state' scenario is soon confronted by uncertainties concerning the geological and hydrological characteristics of the region and basin containing the repository, the proper physical-chemical-geotechnical models governing the flow of water through rock porosities and random fractures, the precise evaluation of dissolution rates that are very sensitive to water chemistry, and the retardation of nuclides by different materials.

Geology appears to be a science in an evolving and unsettled state. There seems to be disagreement among well-qualified workers in geology and the geotechnical sciences and engineering as to underlying theories and models for geological structures, processes, and evaluations of performance, behavior, or function. The issues in repository design and risk evaluation seem to be dependent upon geotechnical questions as to models of, say, water flow in rock and models of geological and hydrological structures that can be inferred from observations and data about a repository site, region, and basin, and the answers to these questions are uncertain and disputed. The uncertainties as to correct models and the uncertainties concerning parameter values within a given model or theory, seem to require interrogation of professionally qualified individuals so as to both estimate/infer models and parameters as well as to identify the attendant levels/kinds of uncertainty that accompany and qualify such estimates. Probabilistic thinking about uncertainties has been gaining increasing acceptance in the geotechnical sciences and engineering.

B. Expert/Professional Opinion

Given that my own background is in statistics, information theory, and the foundations of probability and decision making and not in the geotechnical areas, I have thought about the characterization and evaluation of the information available from the geotechical area and its participants rather than about the geotechnical questions themselves. The area of concern to me is known as the area of 'expert opinion.' The viewpoint is that risk or safety analyses for complex systems operating in complex and poorly understood geological environments can only be carried out through extensive reliance on information gleaned from 'experts' in the relevant geotechnical areas. This proposal raises the following obvious issues:

(i) identification of those aspects of repository evaluation that are controversial enough or uncertain enough to justify recourse to subjective input;

(ii) identification of experts or professionally qualified individuals in the areas of concern;

(iii) choice of information elicitation framework and approach (what are you trying to learn and how do you inquire into it);

(iv) calibration/validation of the selected individuals;

(v) representation of the elicited information in a form suitable for integration into the evaluation process;

(vi) aggregation of the opinions of experts with each other;

(vii) aggregation of expert opinion with objective data such as site measurement and physical/chemical theory;

(viii) propagation of information and uncertainties through the system;

(ix) utilization of the aggregated information and its inherent uncertainties, imprecisions and indeterminancies so as to generate and display a risk evaluation;

(x) implications of the evaluation methodology and its limitations for the level of regulation and the choice of quantities/aspects to be regulated ('regulatables').

My own expertise is largely in (v)-(ix). I am doubtful that there exists much real knowledge concerning (iii)-(ix). Issues (i), (ii) do not seem to me to be problematic. Issue (x), the one of greatest concern to NRC, seems to me to be best dealt with from a technical point of view only after gaining some clarity on the earlier issues. Should NRC be unwilling to take the longer view that my position implies, then perhaps they need to pay more attention to their prospect of success when challenged and forced to submit to peer and judicial review.

C. Outline of Contents

In the remainder of this memorandum I shall comment on (iv) - (ix). My comments will include suggestions for broader forms of probabilistic reasoning that have the advantage over the usual numerical probability representation of uncertainty that they can better express the limits to our knowledge. Risk assessment for environmental risk problems needs to be particularly cautious in that it typically deals with somewhat poorly understood phenomena, possible and improbable events, and severe consequences attendant upon these events. These features compel a rational, cautious approach and militate against a 'philosophy' I have heard expressed at LLL that "we need answers"; the clear implication of such a statement is that answers themselves take precedence over the issues of the validity and meaningfulness of the 'answers'. Such a philosophy makes Bayesian decision analysis attractive for it promises a simple approach and simple answers. Indeed Bayesian analysis easily leads to a trivialization of a hard problem and the perhaps unintentional coverup of this trivialization by a mass of calculations and spurious data.

In Section II we introduce the notions of modal, comparative, and interval-valued probability and roughly indicate their applicability to uncertainty representation. In Section III we touch on the issue of how to determine the expertise of an expert. Section IV addresses the tangled problem of the aggregation of experts' opinions and suggests that it is inadvisable to follow the usual line of seeking a consensus opinion. Section V briefly addresses the issues of the combination of objective and expert data and the use to which the interval-valued representation of the aggregated information can be put to obtain a utility-type risk assessment methodology capable of rating a system design and comparing systems. Section VI provides a brief summary of the position we have presented in this report.

We do not address the problem of how a regulatory agency such as NRC should then use such a risk assessment methodology as a basis for its regulation. Once the methodology is clarified and its limitations understood we think that the selection of regulations will not pose a great difficulty. The selection of 'regulatables' and the level at which one regulates is only a complex problem, from the nonpolitical, technical perspective, when one has only a muddled conception of the prospects and limitations of risk assessment methodologies.

II. Representation of Expert Opinion

A. Forms of Knowledge and the Bayesian Position

The numerous forms of the response of an individual to a question are neither fully catalogued by psychologists nor by linguists. For our purpose we simplify this issue by assuming questions that evoke unambiguous, determinate responses. The paradigm would be a selection by the individual of a single item from a finite list presented to him. A single choice without further qualification could be classified as a deterministic response. For example, geologist G when confronted with site data D and after internal consultation with his theory T may respond by assuming that the site is precisely described by hydrological structure H_0 (ignoring specific model parameter values).

If G is pressed in the elicitation process we may learn that while he is confident that H_0 is correct he can also conceive of hydrologies $\{H_i\}$ that might also conform to D and T. To further develop this illustrative example we assume that G feels that the alternatives $\{H_i\}$ are collectively less likely to be the correct site description than is H_0. The elicitation process might then attempt to further refine this response. In the hands of Bayesian decision analysts G would be 'guided'to produce specific numerical probabilities p_0 for H_0 and p_i for H_i. If G is consistent then presumably $p_0 > \sum_{i>0} p_i$. It is of course a desirable state of affairs to have a precise numerical probability description for the uncertainties in G's knowledge of the true H provided that it reflects something more than wishful or blinkered thinking on the part of the analyst. Unfortunately, I am led to the conviction, shared by others, that a precise numerical probability specification for an individual's uncertainties only rarely has descriptive validity for that individual and is even less often of significance to a decision maker who is not himself the individual being interrogated.

The classical subjectivist/personalist/Bayesian position as initiated by F. Ramsey and B. de Finetti, as refined and made coherent by L. J. Savage, and as explicated by D. Lindley and a host of business-school oriented popularizers, is founded upon real insights. However, these insights are then abused by an insistence that all expressions of the knowledge of a 'rational' individual can be treated identically. I fully agree with the proposition that individuals often possess relevant knowledge about a decision problem that they are unable to explicitly account for. An exaggerated objective theory of decision making (e.g., rigid frequentist-based statistics) improperly excludes such knowledge from an explicit role in the decision process and is thereby in error. However, the subjectivist/personalist position gravely errs when it insists that knowledge must always be expressed in a form suitable for a certain kind of decision making (there is a gap between knowledge and action that the subjectivists bridge too glibly) and that there is only one mode for such expression in terms of numerical probability. (There are technical disputes within the subjectivist community as to the precise nature of this numerical probability, with de Finetti in favor of finitely additive measures rather than the standard countably additive measures. This dispute though has little bearing on the problems likely to be encountered in the Waste Management Program.) One cannot postulate the form knowledge must take any more than one can postulate objective facts about the world. This is an observation so obvious that it would not need making were it not for the influence of the Bayesian decision analysts. One must ascertain something of the nature of the knowledge of a particular kind of individual (e.g., well-qualified hydrologist who is an adherent of a particular school of geological thought) faced with a certain type of question (e.g., location of aquifers in a region with unusually little ground water) and use this to guide the choice of representation for the knowledge thereby elicited.

A second consideration that is given too little weight by the Bayesian decision analysts is the difference between an expert being interrogated as one of a panel of experts to provide information to a decision analyst, who in turn will provide information to a decision maker, and the baseline subjectivist/personalist scenario of a decision maker eliciting himself. The gap between knowledge and action when the knowledge is that of the one who must act is narrower than the gap when knowledge is provided by one party for the use of another. In this latter case, which is the case of interest, there is even less justification to immediately distort knowledge to fit the needs of action.

B. Dissatisfactions with the Standard Probabilistic Framework

Attempts to move away from a deterministic framework have hitherto had little alternative but to move to a framework of numerical probability. The widespread recognition that this numerical framework might be unrealistic has found expression in references to 'uncertainty of uncertainty' and to language requesting ranges for or bounds on probability. Recent articles by D. Bazelon[1979] on judicial review of regulatory practices, T. Page[1978] on assessing environmental risk in a judicial framework, and a talk at LLL on environmental law by Prof. Gary Widman have all emphasized the growing awareness by the judiciary that a defense of regulatory actions will require a showing of an underlying rational methodology and respect for those areas where one is essentially ignorant. The form that disclosure of ignorance can take has been suggested to be intervals of probability. One must perhaps pay more attention to the distrust of unfoundedly precise statements than to the specific advice as to how to produce properly conservative statements.

It is instructive in this connection to recall the Lewis Commission Report[1978] on WASH-1400 that led NRC to withdraw its support of WASH-1400. To quote from this report:

> When there is an inadequate data base, ...the limits of knowledge should be stated, without pressure to quantify (other than bounding) that which is unquantifiable. (p. xi)
>
> In general, avoid use of the probabilistic risk analysis methodology for the determination of absolute risk probabilities for subsystems unless an adequate data base exists and it is possible to quantify the uncertainties. (p. xi)
>
> RSS [WASH-1400] has a nearly universal practice of fitting every piece of data it has to a log-normal distribution, and, whenever an entirely unknown distribution needs to be subjectively chosen, to choose a log-normal distribution. ... So within the errors listed in RSS, we accept the log-normal as an acceptable summary of most data. (p. 9)
>
> It is our view that use of subjective probabilities is necessary and appropriate ... but their use must be clearly identified and their limits of validity must be defined. (p. 9)

> The only situation in which one might be concerned about a subjective probability leading to really major errors is where there is no experience at all. (p. 10)

> The choice of one model over another generates an uncertainty, but within that uncertainty the use of the model is justified, provided the uncertainty is estimated and indicated. (p. 10)

> There are cases in which an entire distribution is "derived" from only a few data points. ... But the uncertainties associated with using it are then correspondingly large, and need to be taken into account and propagated through the entire calculation. (p. 11)

The Lewis Commission Report, critical as it is of RSS or WASH-1400, is in my view, still too optimistic about the use and control of subjective input and the severity of the uncertainties generated when there is informed dispute about the choice of model. The distinction between choice of model and choice of parameter within a model is not an intrinsic mathematical issue since the set of models can itself be parameterized. However, discussions on this issue usually reflects a qualitative vs quantitative distinction. When there is debate about, say, the physical mechanism for flow of water through rock and this leads to choices between say Darcian flow and fracture flow, then one is in an environment in which it will be much harder to characterize uncertainties than when one is in doubt about a typical parameter value.

The Lewis Commission acceptance of subjective input seems to presume an agreed upon and well-substantiated methodology for eliciting and validating subjective probabilities. They refer to the predictive accuracy of horse race bettors and we could supplement this by reference to studies of the accuracy of weather forecasters. However, there is reason to believe that these validation studies are uninformative about the performance of 'experts' in the areas of concern to nuclear risk assessment where there is far less data and experience and a much greater reliance upon theory. The Lewis Commission may have been insufficently skeptical about subjective input possibly because it was insufficiently knowledgeable in this area. I suspect on the basis of my brief exposure to the areas of Waste Management and Seismic Safety Design at LLL that the complexity of nuclear systems and their interaction with a poorly understood geological environment and the unique nature of many components means that significant data will have to come from 'qualified' individuals who are yet poorly informed about the subjects of the inquiries. Representing the information gained from such experts and the uncertainties in the information is a problem that still requires fundamental research. Facile ad hoc approaches are unlikely to bear-up under the scrutiny of judicial or peer review.

The basic Bayesian approach, wherein one elicits a response either directly in terms of probabilities or implicitly in terms of probabilities through binary comparisons of likelihood, is both epistemologically hazardous and likely to produce numbers reflecting little more than the happenstance personal interaction between expert and analyst. Nor does recourse to sensitivity analysis, the Bayesian security blanket, provide an answer here. While it is of interest to know the extent to which variations in the arguments of a function change the value of the function, this information says nothing about the validity of any particular assignment of values to the arguments of the function. Sensitivity analysis is not a guide to or guarantor of truth. Further elicitation is the Bayesian's only recourse when there is sensitivity to an elicited parameter, and this is just more of the same medicine.

C. Extended Forms of Probabilistic Reasoning

1. Possibly

One needs to employ a variety of probabilistic concepts in characterizing expert opinion. This variety is little discussed in the probabilistic or philosophical literature but can be examined in Walley and Fine[1979] and in Walley[1979] and is introduced in this report. The concepts range from 'possibility' to 'probability' (modal) to 'relative probability (comparative)' to 'interval-valued probability' to 'statistical hypotheses' and thence to the usual 'numerical probability.' The least refined notion is that of possibility, which we can denote by 'MA' read 'A is possible.' The modality of possibility has been studied in the discipline of philosophical logic and it is well-described in Hughes and Cresswell.[1972] Possibility has many possible meanings or interpretations of which examples are logical, physical, ethical, and practical. It is logically possible for a ball released from my hand to land on the moon but not physically possible. It is physically possible for a tossed dime to land on edge but not practically possible. (Ethical possibility is of little interest in this connection but might be of some interest if we treated the range of decision rules.) The notion of possibility is reflected in control theory in the term 'bounded uncertainty.' Bounded uncertainty refers to a state of knowledge about, say, a parameter θ that it lies in a set Θ and that we know little about which subset of Θ contains θ.

Elsewhere we have used the term 'indeterminate' to refer to phenomena in which our knowledge amounts to an ungraded list of possibilities. Characterizations of knowledge exclusively in terms of possibility seem to fit only with a minimax design philosophy. Referring to repository design problems, we suspect that we can do better than by exclusively relying on the notion of bounded uncertainty, but this notion might be the only realistic one for particular aspects about which we know very little (e.g., possible future human intrusion into the repository.) We do not emphasize this weak form of probabilistic reasoning here, although we note that decision making based upon a notion of possibility is also contemplated in the recent literature on fuzzy sets (Zadeh [1978], Yager [1979]).

2. Probably

A somewhat stronger notion than that of 'possibility' is provided by the notion of 'probable' or 'not improbable.' We denote 'A is not improbable' by 'PA.' Clearly false MA implies false PA, and that is about the extent of the relationship between possibility and probability in this basic setting. 'PA' then requires that A be possible and is a further distinction or refinement among the class of possible events. Elementary axioms for PA, when we assume an event collection (algebra) $\mathcal{A}$ of subsets A, B, C, .. of a sample space Ω and we let $\emptyset$ denote the complement Ω^c of Ω or the impossible event, are as follows:

M1. $P\Omega$ - the sure event is not improbable.

M2. PA or PA^c - either A or its complement is not improbable.

M3. PA and $B \supset A$ implies PB - if A is not improbable and A implies B, then B is not improbable.

Clearly P is a rather primitive notion having seemingly no quantitative structure. In fact, we can develop a rough numerical structure for the qualitative notion 'probably' through the device of 'almost uniform partitions.' Select a partition $\{S_i\} = \{S_1,\ldots,S_n\}$ of Ω where $S_i \varepsilon \mathcal{A}$ and such that if α is any subset of $\{1,\ldots,n\}$ of size $\left\lfloor\frac{n}{2}\right\rfloor +1$ and β any subset of size $\left\lfloor\frac{n}{2}\right\rfloor -1$, then $P\left(\bigcup_{j\varepsilon\alpha} S_j\right)$ and false that $P\left(\bigcup_{j\varepsilon\beta} S_j\right)$. Hence $\{S_i\}$ is such that the union of more then one-half of the events in the partition is always not improbable and the union of less than half of them is improbable. There always exists such a partition since any two-fold partition of Ω will satisfy the above conditions.

Given any event $A \varepsilon \mathcal{A}$ we can then assign a numerical interval $(\underline{d}, \overline{d})$ to A relative to $\{S_i\}$ through

$$\underline{d}\ (A) = \text{MAX} \left\{\frac{k}{n}\colon \alpha \subseteq \{1,\ldots,n\},\ \bigcup_{j\varepsilon\alpha} S_j \subseteq A\ ,\ k = ||\alpha||\right\},$$

$$\overline{d}\ (A) = \text{MIN} \left\{\frac{k}{n}\colon \bigcup_{j\varepsilon\alpha} S_j \supseteq A,\ k = ||\alpha||\right\}.$$

There are various technical issues here, including the dependence of $(\underline{d}, \overline{d})$ upon the particular partition $\{S_i\}$, but we introduce this assignment for illustrative purposes. It is of interest to note that if we identify $\underline{d}$ with $\underline{P}$ and $\overline{d}$ with P, then $\underline{d}$, $\overline{d}$ satisfy the IVP axioms to be introduced in Subsection 4.

A somewhat different approach to associating a numerical structure with P would be through the agency of an agreeing probability measure. We say that the probability measure μ agrees with or represents P if there is some threshold t independent of A such that

$$PA \text{ if and only if } \mu(A) \geq t.$$

The basic case would be one in which PA iff false PA^c, and then we would have to take $t = 1/2$ and deal with the case of $\mu(A) = 1/2$. Should there exist agreeing μ then one would look at the family M_p of measures agreeing with P and define

$$\underline{d}(A) = \inf \{\mu(A) : \mu \varepsilon M_p\}, \quad \overline{d}(A) = \sup \{\mu(A); \mu \varepsilon M_p\}.$$

The difficulty with this natural approach is that M_p can be empty--there need not exist any agreeing μ for P. In other words, 'not improbable' does not necessarily have its origin in any underlying numerical probability. However, the qualitative, weak concept of 'not improbable' can support an interval-valued numerical probability assignment to events.

3. Comparative Probability

We now turn to the more substantial concept of comparative probability (CP). By CP we refer to an ordering or ranking of the 'likelihood' of events denoted by '$A \geq B$' and read 'A is at least as probable as B.' In terms of $\geq$ we can define $>$, $\sim$ as follows:

$$A > B \text{ iff } A \geq B \text{ and not } B \geq A,$$

$$A \sim B \text{ iff } A \gtrsim B \text{ and } B \geq A.$$

We read '$A > B$' as 'A is more probable than B' and '$A \sim B$' as 'A and B are neither more probable than the other.' The familiar axioms for CP are (see Fine,[1973] Kaplan and Fine):[1977]

CP1. (Complete) $A \geq B$ or $B \geq A$.

CP2. (Transitive) $A \geq B$ and $B \geq C$ imply $A \geq C$.

CP3. (Nontrivial) $\Omega > \emptyset$.

CP4. (Positivity) $A \geq \emptyset$.

CP5. (Cancellation) $A \geq B$ iff $A-B \geq B-A$.

This axioms for CP have been presented in Walley and Fine[1979] and lead to weaker notions of CP in which transitivity is restricted or the ordering is allowed to be partial. These refinements, though, need not concern us here.

CP provides a natural representation for expert opinion in that it conforms to the basic device of binary comparisons that is central to the Von Neuman-Morgenstern approach to utility theory and Savage's approach to subjective probability, and both of these approaches are central to, say, Bayesian decision analysis. While some Bayesian decision analysts, in their haste to reach a numerical conclusion, attempt to directly elicit numerical probabilities for events, the fundamental approaches to subjective probability are based upon eliciting binary comparisons of likelihood from which one then attempts to construct a numerical probability representation.

We say that the probability measure μ agrees with or represents the CP relation $\gtrsim$ if for all A, B in $\mathcal{A}$,

$$A \gtrsim B \text{ iff } \mu(A) \geq \mu(B).$$

One asks the expert to compare pairs of events as to their relative likelihood and then searches for an agreeing μ. The agreeing μ will generally not be unique but the range of possibilities can be narrowed by augmenting the event space $\mathcal{A}$ to, say, one in which we have adjoined an independent side experiment consisting of N tosses of a fair coin; one then elicits further event comparisons (e.g., is $A \in \mathcal{A}$ more or less likely than at least k heads in N tosses?).

Let $M_{\geq}$ denote the family of probability measures agreeing with $\geq$. Then $M_{\geq}$ can be empty and this implies that the ordering $\geq$ does not agree with any probability measure. Restated, it need not be possible to construe a CP ordering as an approximate statement of numerical probability. Hence we see that CP is not just an approximation to numerical probability but in fact an independent and more general notion of probabilistic reasoning. The simplest examples of CP orderings in which $M_{\geq}$ is empty require that Ω have five atoms and $\mathcal{A}$ then contains the $2^5 = 32$ subsets of Ω.

For example, the geologist G might assert that on the basis of his experience and site data there are five possible hydrological structures $H_1, .., H_5$ that might describe the site. Elicitation might proceed by asking G to compare the likelihood of, say, H_i to H_j. Perhaps on the basis of his responses we learn that

$$\emptyset < H_1 < H_2 < H_3 < H_4 < H_5;$$

i.e., all five hydrologies are possible but hydrology H_{i+1} is more likely to be correct than hydrology H_i. We could then refine this ordering, if G were willing, by moving to such more complex binary comparisons as, "Is H_5 more or less likely than the possibility that the hydrology is either H_3 or H_4?" If G answers that "the hydrology is more likely to be either H_3 or H_4 than it is to be H_5" than we record this as

$$"H_5 < H_3 \cup H_4",$$

which we will write more compactly as "5 < 34".

If G could respond to all of the finitely many binary comparisons then the record of his responses could come out as (where for convenience 'ij' denotes 'H_i or H_j')

$$\emptyset < 1 < 2 < 3 < 12 < 13 < 4 < 14 < 23 < 5 < 123 < 24 < 34 < 15 < 124$$
$$< 25 < 134 < 35 < 234 < 125 < 135 < 45 < 1234 < 145 < 235$$
$$< 1235 < 245 < 345 < 1245 < 1345 < 2345 < 12345 = \Omega.$$

Consideration of the comparisons

$$13 < 4,\ 14 < 23,\ 34 < 15,\ 25 < 134, \tag{*}$$

reveals that $M_{\geq}$ is empty for this ordering. If G believes (*) to be correct, even after reconsideration, then G's knowledge about the hydrology of the site is not representable by numerical probability. A Bayesian decision analyst faced with G would have to induce him to change his mind. In a CP setting, however, we can accept and work with this set of responses.

A publicly defensible introduction of expert/professional opinion in the environmental risk area and particularly in the nuclear area should be based upon a CP representation of knowledge, where the knowledge so represented should include not only an assessment of the likelihood of the events of interest but also approximate gradings of the degree of conviction or validity of the particular response. Although this has not been attempted elsewhere to the best of my knowledge, I would recommend that elicitation of expert opinion in the nuclear waste repository assessment problem, at least as regards structural, theoretical, or model selection features (i.e., not necessarily as regards assessment of numerical parameters within a specific model) be on the basis of responses to binary comparisons with an accompanying coarse grading of the reliability of the particular response.

Let $(S,\mathcal{S})$ denote the pair of a scaling sample space S and its power set (set of all subsets) $\mathcal{S}$. Let the ordering $\geq_{\mathcal{S}}$ on $\mathcal{S}$ be almost uniform in the sense that if $||A||$ denotes the cardinality of A then

$$||A|| > ||B|| \text{ implies } A >_{\mathcal{S}} B.$$

The CP- space $(S, \mathcal{S}, \geq_{\mathcal{S}})$, if it is unlinked to the space $(\Omega, \mathcal{A}, \geq_G)$ of particular interest (e.g., the set of hydrologies and their ranking) can then serve to rate the reliability of a particular response through recourse to the product space $S \times \Omega$, the product event algebra $\mathcal{S} \times \mathcal{A}$, and an independent joint order $\geq$ on $\mathcal{S} \times \mathcal{A}$ where characteristic properties of an independent joint order include:

(i) $\geq$ is a CP ordering on $\mathcal{S} \times \mathcal{A}$;

(ii) $(\forall A, B \in \mathcal{A})$ $S \times A \geq S \times B$ iff $A \geq_G B$;

(iii) $(\forall T, R \in \mathcal{S})$ $T \times \Omega \geq R \times \Omega$ iff $T \geq_{\mathcal{S}} R$;

(iv) $(\forall A, B \in \mathcal{A}, T, R \in \mathcal{S})$ $A >_G B$, $T \geq_{\mathcal{S}} R$ implies $T \times A \geq R \times B$;

(v) If in addition to (iv) either $A >_G B$ and $T >_{\mathcal{S}} \phi$ or $A \geq_G \phi$ and $T >_{\mathcal{S}} R$ then $T \times A > R \times B$.

Properties (ii), (iii) just assert that the joint order has the correct marginal orderings. Properties (iv), (v), reflect a core property of independence between the marginals. Property (i), while obvious, is in fact not always satisfiable within the version of CP whose axioms we have presented earlier. There can exist marginal CP orderings such that there is no joint ordering satisfying (ii) and (iii) let alone (iv), (v). To ensure the uniform existence of a joint CP ordering we would have to enlarge the set of CP orderings by weakening the CP axioms. Some suitable possibilities are suggested in Walley[1979] and Walley and Fine.[1979]

Ignoring the technicalities for the present, we have proposed a rather simple, once understood, and direct scheme for the representation of expert/professional opinion which incorporates constraints of rationality (e.g., transitivity and cancellation) and does not place an undue burden of precision on the expert to announce results that he may not possess. We also are able to simultaneously record some rating of the reliability of this response. Reliability information is, in my view, quite essential for keeping track of the state of our knowledge and for the subsequent process of rational aggregation of subjective inputs and for aggregation with such objective inputs as site measurements.

4. Interval-Valued Probability

By interval-valued probability (IVP) we refer to a concept in which we assign an interval with lower endpoint $\underline{P}(A)$, called the lower probability, and upper endpoint $\overline{P}(A)$, called the upper probability, to an event A. Roughly the idea is that the interval $(\underline{P}(A), \overline{P}(A))$ represents not only our best guess as to 'the' probability of A but also the range of our uncertainty or lack of knowledge about this probability. More fundamentally we should not assume that there is an underlying 'true' probability μ (A) contained in the interval. Indeed the basic axioms for IVP are compatible with the nonexistence of any underlying numerical probability measure μ and this claim will be made precise below. There are several approaches to and axiomatizations for IVP[2,4,5,10,11,12,14] and we present an axiomatization due to P. Walley.

IVP1. $\underline{P}(A) \geq 0$, $\underline{P}(\Omega) = 1$.

IVP2. If $A \cap B = \phi$ then

$\underline{P}(A) + \underline{P}(B) \leq \underline{P}(A \cup B)$ (super additivity),

$\overline{P}(A) + \overline{P}(B) \geq \overline{P}(A \cup B)$ (subadditivity).

IVP3. $\overline{P}(A) = 1 - \underline{P}(A^c)$.

Elementary consequences of these axioms include:

T0. $\overline{P}(A) \geq 0$, $\overline{P}(\Omega) = 1$, $\underline{P}(\phi) = \overline{P}(\phi) = 0$.

T1. $(\forall A)\ \overline{P}(A) \geq \underline{P}(A)$

T2. If $A \supset B$ then $\overline{P}(A) \geq \overline{P}(B)$, $\underline{P}(A) \geq \underline{P}(B)$.

T3. If $A \cap B = \phi$ then

$$\underline{P}(A \cup B) \leq \underline{P}(A) + \overline{P}(B) \leq \overline{P}(A \cup B).$$

It is of interest to inquire into the relationship between IVP and numerical probability (NP).

<u>Def</u>. We say that $\underline{P}$ is dominated by a probability measure μ, $\underline{P} << \mu$, if $(\forall A \varepsilon \mathcal{A})\ \mu(A) \geq \underline{P}(A)$.

If $\mu >> \underline{P}$ then it is easy to see that $\overline{P} >> \mu$ and $(\forall A)\ \overline{P}(A) \geq \mu(A) \geq \underline{P}(A)$.

T4. There exist $\underline{P}$, $\overline{P}$ satisfying IVP1-3 such that they are undominated.

Even if $\underline{P}$ is dominated by some μ it need not be determined by μ. The class of IVP determined by probability measures are the so-called lower envelopes. Let

$M_{\underline{P}} = \{\mu: \mu >> \underline{P}\}$

Then $\underline{P}$ is a lower envelope iff

$(\forall A)\ \underline{P}(A) = \inf\ \{\mu(A) : \mu \varepsilon M_{\underline{P}}\}$.

Any family M of probability measures can induce a lower envelope through

$\underline{P}(A) = \inf\ \{\mu(A): \mu \varepsilon M\}$.

In this case $M \subseteq M_{\underline{P}}$. One can further refine the relationship of IVP to NP by introducing degrees of regularity extending up to the belief functions discussed in G. Shafer,[1976] but the above is sufficiently illustrative for the purposes of this memorandum.

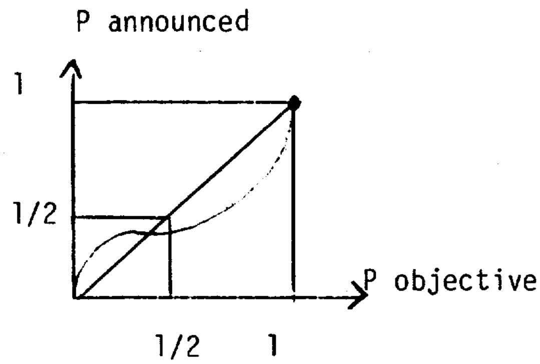
We can relate IVP to CP through

T5. If $\underline{>}$ satisfies CP1-4 then there exists $\underline{P}$ satisfying IVP1-3 such that

$A \underline{>} B$ iff $\underline{P}(A) \underline{>} \underline{P}(B)$ and $\overline{P}(A) \underline{>} \overline{P}(B)$.

In fact T5 holds even if we restrict $\underline{P}$ to be the belief function (monotone of order infinity IVP) studied by Shafer and by A. Dempster. The import of T5 for us is that it enables us to connect the epistemologically sound CP representation of expert opinion with the computationally more tractible framework of IVP. Our proposal would be to represent expert knowledge by CP, as discussed in Subsection C.3, and then transform CP into IVP. This latter transformation raises questions that have to be addressed since there are many IVPs that can represent a given CP.

We close this section by observing that we are unconvinced of the soundness of a process of direct elicitation of IVP. It is true that some of the work on a subjective basis for IVP (e.g., [4,5,11]) requires a as well as the way subjective probability is a distortion of objective probability in those situations where ample data on a repeatable experiment is available to the individual being elicited. A typical distortion characteristic is:



In the above characteristic an objective small probability is over-estimated while a large probability is underestimated. The kinds of events being assessed seem to determine whether probabilities will be over or under estimated. In all cases though the distortion is most pronounced at the extremes and unfortunately one expects that much expert elicitation will in fact concern extreme cases of low probabilities. Nonetheless, the results of psychological research on subjective probability distortion does provide some guide for correcting the expert reports.

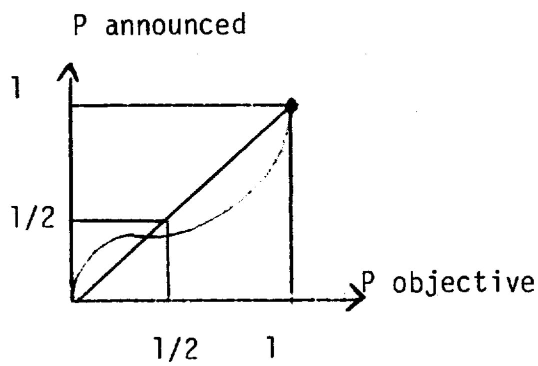
We should also note that the subjective probabilities that are elicited are dependent upon the elicitation process. Probability elicitation is not a neutral process. Bayesian decision analysts typically agree with this but they claim a personal ability to minimize such effects.

III. Calibration/Validation of Expert/Professional Opinion

The issues here become apparent when one views the 'expert' as a measuring instrument with verbal responses to queries constituting the result of the measurement on the quantity being queried. Clearly one needs to ascertain the accuracy and precision of this complex instrument connected to the subject of inquiry in a poorly understood fashion. The 'expert' instrument needs to be validated insofar as we can determine its relation to truth as well as calibrated to account for systematic/persistent distortion and biases and sensitivity to the form of interrogation.

The validation issue is hard to deal with in contexts such as those that can occur in the nuclear risk assessment area where there is no record of past performances of the expert on the same question. In this case one infers validity from professional qualifications (e.g., training and standing among the peers of his/her profession) as well as from a record of performance on 'similar' questions. Of course, what constitutes 'similarity' is itself a substantial issue, but not perhaps one that needs careful explication here. For example, one might test the presumed expert by letting him/her examine some of the data on a site that has been carefully explored but with which the expert claims no familiarity. One can then compare the expert's judgments to the objectively known situation. In this way one learns not only about the individual expert but perhaps also about the capabilities of the field of which he is a member.

Studies in the subjective probability literature (e.g., A. Tversky and D. Kahneman[1974]) reveal that individuals tend to predictably distort their experience in announcing probability assessments. Tversky has explored several phenomena that tend to produce assessment errors (e.g., anchoring about initial guesses or suggestions, emphasis on the easily recalled as being more likely) as well as the way subjective probability is a distortion of objective probability in those situations where ample data on a repeatable experiment is available to the individual being elicited. A typical distortion characteristic is:



In the above characteristic an objective small probability is over-estimated while a large probability is underestimated. The kinds of events being assessed seem to determine whether probabilities will be over or under estimated. In all cases though the distortion is most pronounced at the extremes and unfortunately one expects that much expert elicitation will in fact concern extreme cases of low probabilities. Nonetheless, the results of psychological research on subjective probability distortion does provide some guide for correcting the expert reports.

We should also note that the subjective probabilities that are elicited are dependent upon the elicitation process. Probability elicitation is not a neutral process. Bayesian decision analysts typically agree with this but they claim a personal ability to minimize such effects.

IV. Aggregation of Expert Opinions

The issue of the aggregation of expert opinion concerns the derivation of a single representation for the somewhat divergent opinions of a group of experts queried about the same or logically related issues. The usual approaches assume a numerical probability representation for the individual expert opinions and then strive towards a consensus probability distribution. There are two main avenues of approach:

(i) calculation of a consensus distribution by a decision analyst/statistician who is likely not himself an expert in the area being queried;

(ii) generation of a consensus by the group itself, either on the basis of controlled communication between the members (e.g., Delphi) or on the basis of simultaneous face-to-face interaction.

I would contend that the urge towards the generation of a consensus probability distribution is inappropriate in the area of environmental risk assessment and particularly in the nuclear area. One must exercise care to qualify in an explicit fashion the limits to one's knowledge, and divergent expert opinion is _prima facie_ indicative of a limit to knowledge. A representation of divergence of qualified opinion is necessary. Divergence should not be completely suppressed as would occur in the generation of a consensus distribution. Interval-valued representations for expert opinion have the advantage that they can be aggregated to form an interval-valued 'consensus' in which the interval widths in the consensus can reflect the indeterminacy inherent in the expert opinion.

The issue of the generation of an interval-valued consensus is at present still a research issue. It is also true that the generation of consensus probability distributions is still a research issue. For example, Delphi has been soundly criticized in a recent RAND Report as having no relation to improved validity, inconclusive studies of the properties of consensus processes appear in the recent literature, and no one has been successful in dealing with overlap and independence of experts.

The aggregation question cannot be answered just by examination of the responses of the individual experts. The mathematical analogy here is to the formation of a joint probability distribution from given marginal probability distributions. The marginals do not determine the joint distribution, one requires additional information about the linkage between the marginals. Such information could be that for physical reasons the marginals are unlinked, and thus we model the joint distribution as a product of marginals. Should the information be that the marginals are linked or dependent then one would need to know the form of the dependence and this could become quite complex. Knowledge of a correlation, say, between two variables would not suffice to determine a joint distribution even if the marginals were all Gaussian. There is no reason to expect the aggregation of expert opinion to be a simpler problem than the aggregation of probability marginals.

The above is not to say that we are without resources. Various aggregation rules suggest themselves such as the Dempster rule of combination and the following highly conservative aggregation rule. If $\underline{P}_i$ is the lower probability describing the knowledge of the ith expert then

$$\underline{P}^*(A) = \min_i \underline{P}_i(A) \quad , \overline{P}^*(A) = \max_i \overline{P}_i(A)$$

are new lower and upper probabilities that describe the group of experts. A somewhat less conservative rule might have us eliminate outlying expert opinion before aggregating. One needs to examine the properties of such rules so as to build towards an intuition or methodology capable of selecting a reasonable rule when faced with a particular panel of experts. Which rule one uses will depend upon the experts' field (e.g., geology, hydrology, corrosion, solubility of nuclides, structural engineering) as well as the training, background, and theoretical dispositions of the experts (are they all adherents of the same theory or do they cluster into schools of thought?).

My recommendation is that research needs to be sponsored but that if I had to aggregate within a school of thought I would at present incline towards a lightly censored conservative rule of the form: select a threshold λ,

$$\hat{\underline{P}}\ (A) = \underline{P}_i(A) \text{ if } \underline{P}_i(A) \text{ is the } \lambda^{th} \text{ largest of } \{\underline{P}_j(A)\}.$$

The derivation of well-founded aggregation or combination rules seems to me to require a canonical interpretation of experts so that they can all be placed upon the same footing preparatory to combination. For example, when there is a salient likelihood function in the problem we may be able to interpret an expert as corresponding to an effective sample size and the so-called vacuous initial lower probability. One can then aggregate experts according to their individual sample sizes with dependence between experts being modeled as sample overlap.

V. Using the Aggregated Expert Opinion

A. Combination With Objective Data

Objective and frequency-of-occurrence data can be given an interval-valued representation following the technique of M. Wolfenson.[1979] Once, say, measurement data have been so represented they become yet another 'expert' and aggregation of data with experts can proceed as is done for aggregation of experts themselves. However, ideas on this subject need to be regarded cautiously as there is a near absence of research and working experience.

B. Rating Actions

Once one has the aggregated expert opinion and measurement data represented by IVP one then propagates the IVP representation for parameter and model uncertainties through the system to arrive at an IVP description for the output states of interest. For example, one may calculate $(\underline{P}(A), \overline{P}(A))$ where A is the event that in year n after sealing of the repository at least C Curies of nuclides will be deposited on the surface. The particular interval will be dependent upon the repository design δ incorporating site selection, construction design, backfilling techniques, and choices of canister materials and waste forms. Hence we would have a collection $\{\underline{P}_{\delta}, \delta \varepsilon D\}$ where D is the set of design alternatives. The event algebra $\mathcal{A}$ might contain the events of the form described above relating to the time history of the emissions of radioactive material into the biosphere.

One then wishes to choose a good design from D, or, from the viewpoint of NRC, to establish risk levels representative of small 'harm to man' and calculable from the design δ. The assessment of risk will require the assessment of the degree of harm to man resulting from a given emission history. While I believe that it is no more realistic to assume that one can supply a Von Neuman-Morgenstern utility function to numerically represent 'harm to man' than one can come up with precise numerical probabilities for emission histories, let me assume for the moment that some utility function can be chosen. Perhaps $\mathcal{U}$ is defined on the set $\Omega = \{\omega\}$ in which ω is in fact a sequence $\omega = \omega^1 \omega^2 \ldots \omega^N$ where ω^i is the Curie emission reaching the biosphere in the ith year following sealing of the repository. More simply ω might just evaluate the harm to man due to emissions of a given Curie level in a fixed time period. In any event $\mathcal{U}: \Omega \to R^1$ such that $\mathcal{U}(\omega) > \mathcal{U}(\omega')$ implies that history ω is less harmful than is history ω'.

The system design δ would then be evaluated in the classical Bayesian framework by evaluating $E_\delta \mathcal{U}$, where E_δ represents expectation taken according to the probability measure μ_δ on the output space Ω arising from propagation of uncertainties through system δ. In place of μ_δ we have the IVP $(\underline{P}_\delta, \overline{P}_\delta)$. We can, however, parallel the Bayesian theory by following Dempster[1967] and introducing upper $(\overline{E}_\delta)$ and lower $(\underline{E}_\delta)$ expectations. Bypassing some technical issues of measurability and continuity we introduce the cumulative distribution functions

$$\underline{F}_\delta(x) = \underline{P}_\delta\ (\{\omega;\ u(\omega) \leq x\}),\ \overline{F}_\delta(x) = \overline{P}_\delta\ (\{\omega;\ u(\omega) \leq x\}).$$

It is easily verified that $\underline{F}_\delta$, $\overline{F}_\delta$, are both ordinary cdf's.

We then define

$$\overline{E}_\delta = \int_{-\infty}^{0} F_\delta(x)dx + \int_{0}^{\infty} (1-\underline{F}_\delta(x))\ dx,$$

$$\underline{E}_\delta = \int_{-\infty}^{0} \underline{F}_\delta(x)dx + \int_{0}^{\infty} (1-F\ (x))\ dx,$$

The above definition of $\underline{E}$, $\overline{E}$ can be justified to some extent although we do not do so here.

Hence to each system δ we associate the interval of expectations $(\underline{E}_\delta, \overline{E}_\delta \mathcal{U})$. We can now compare two systems δ, δ' by comparing their associated utility intervals. If, say, $\underline{E}_\delta\, \mathcal{U} > \overline{E}_{\delta'} \mathcal{U}$, then system δ is preferred to system δ'. Some cases of overlap between the intervals may not be clearly resolvable. When this happens we are unable to compare δ to δ' to decide which is better. This may not be a problem though from the NRC viewpoint. NRC might just wish to regulate by setting a minimum acceptable lower expectation.

VI. Summary of Conclusions and Recommendations

The objective of this report has been to critically discuss, and make recommendations concerning aspects of decision-making and modeling that bear on the necessary use of expert opinion in the design and evaluation of nuclear waste repositories.

Briefly, we are doubtful as to the prospects for complete reliance upon the subjectively-based numerical probabilities that lie at the core of Bayesian decision analysis. However, we agree with the Bayesians that a deterministic analysis replacing uncertainties by 'certainty equivalents' cannot be defended either in a rational public forum or in a sophisticated technical forum. We also agree that rational design must incorporate the 'pre-scientific' opinions and beliefs of qualified individuals, especially in an area lacking in objective data and widely accepted scientific models.

Our recommendations concerning the modeling and representation of the opinions and beliefs of qualified individuals (experts) about the uncertainties encountered in repository assessment and design involve the use of the unfamiliar concepts of modal, comparative, and interval-valued (upper and lower) probabilities. These concepts, drawn from the current literature on the foundations of probability and statistics, are selected to provide us with an improved ability to faithfully portray the absence of precise probabilistic knowledge of the kind required to assess the usual numerical probability concept. It seems clear that there are uncertainties affecting repository performance concerning which there is little data (e.g., detailed site hydrological characteristics), debate over proper theoretical models (e.g., flow in rock), and even the most qualified individuals will not know much about the issues in question.

One needs a formal, rational methodology that can incorporate what is known and believed, without over-rating the quality of such beliefs. Comparative and interval-valued probabilities are recommended in this report as filling this need.

Comparative probability provides a direct representation for the results of the basic process of expert elicitation via binary comparisons of likelihood--a device we share with the Bayesians and one also common in utility assessments. Interval-valued probabilities can be used to represent comparative probabilities in a form more suited to computation, including eventual aggregation of expert opinion. While there is also the possibility of direct elicitation of interval-valued probabilities, we are skeptical as to the validity of such a process.

We touch on the issue of validating expert opinion but have nothing to add beyond the caveats discerned in the psychological literature and particularly in the work of Paul Slovic and Amos Tversky.

The issue of the aggregation or amalgamation of the opinions of several experts is a key one and one insufficiently explored in the literature. Our recommendation here is that care be taken to preserve some of the discrepancies in the collective expert opinion. The literature on aggregation is generally aimed at the formation of a consensus opinion. We feel, however, that one must act conservatively in as socially sensitive an area as that of nuclear waste disposal and this suggests explicitly keeping track of the limits to our knowledge revealed by the disagreements between experts. This position is in accord with trends seen in judicial review of regulatory agency actions.

We close in Section V with a mathematical sketch of the way in which comparative and interval-valued models of uncertainty can enter into the repository performance assessment process in a manner parallel to the use of expected utility to represent preferences. Utility theory is, of course, the basis of most statistical decision-making.

It should be understood that much less is known about the mathematical aspects of modal, comparative, and interval-valued probabilities than about the mathematical structure of numerical probability. Furthermore, there is far less working experience with these new concepts than with the familiar numerical concept. Hence one should proceed to use the approach outlined in this report with circumspection, but one should consider this approach for it holds promise for publicly defensible rational probabilistic models of repository performance, models incorporating the best of our subjective and objective knowledge.

VII. Acknowledgments

This work was supported by the Nuclear Regulatory Commission Waste Management Program at the Lawrence Livermore Laboratory, and my participation in this program was due to Dr. Guy C. Corynen. Dr. Corynen guided me to sources of information, answered my questions, and shared with me his perspective as a system theorist on the repository design problem.

References

1. D. Bazelon 1979 , Risk and responsibility, Science, 205, 277-280. (Based upon a talk given at the Amer. Bar Assn. Nat. Inst. on Law, Science, and Technology in Health Risk Regulations, 20 April 1979.)

2. A. Dempster [1967], Upper and lower probabilities induced by a multivalued mapping, Ann. Math. Statistics, 38, 325-339.

3. T. L. Fine [1973], Theories of Probability, Academic Press, New York, Ch. II.

4. R. Giles [1976], A logic for subjective belief, in W. Harper, C. Hooker, eds., Foundations of Probability Theory, Statistical Inference, and Stat. Theories of Science, Vol. I, D. Reidel Pub. Co., Dordrecht, Holland, 41-72.

5. I. J. Good [1962], Subjective probability as the measure of a non-measurable set, in E. Nagel, P. Suppes, A. Tversky, eds., [1960], Logic, Methodology and Philosophy of Science, Stanford University Press, Stanford.

6. G. Hughes, M. Cresswell [1972], An Introduction to Modal Logic, University Paperbacks (reprint with corrections of 1968 book published by Methuen and Co. Ltd., London).

7. M. Kaplan, T. L. Fine [1977], Joint orders in comparative probability, Ann. Probability, 5, 161-179.

8. H. W. Lewis et al., [1978], Risk Assessment Review Group Report to the Nuclear Regulatory Commission, Nat. Tech. Inf. Svce, Springfield, Virginia.

9. T. Page [1976], A generic view of toxic chemicals and similar risks, Ecology Law Quarterly, 7, 207-244.

References (con't)

10. G. Shafer [1976], A Mathematical Theory of Evidence, Princeton University Press, Princeton.

11. C. Smith [1961], Consistency in statistical inference and decision, J. Roy. Stat. Soc., B23, 1-25.

12. P. Suppes [1974], The measurement of belief, J. Roy. Stat. Soc., B36, 160-175.

13. A. Tversky, D. Kahenman [1974], Judgment under uncertainty: heuristics and biases, Science, 185, 1124-1131.

14. P. Walley [1979], Varieties of Modal and Comparative Probability, Ph.D. Thesis, Cornell University, Ithaca, New York.

15. P. Walley, T. L. Fine [1979], Varieties of Modal (Classificatory) and Comparative Probability, Synthese, to appear July.

16. M. Wolfenson [1979], Inference and Decision Making Based on Interval-Valued Probability, Ph.D. thesis, Cornell University, Ithaca, New York.

17. R. Yager [1979], Possibilistic decision making, IEEE Trans. on Systems, Man, and Cybernetics, SMC-9, 388-392.

18. L. Zadeh [1978], Fuzzy sets as a basis for a theory of possibility, Fuzzy Sets and Systems, 1, 3-28.